TeraGrid Overview and Life Science Gateway

Rick Stevens

University of Chicago and Argonne National Laboratory

Feb 2006

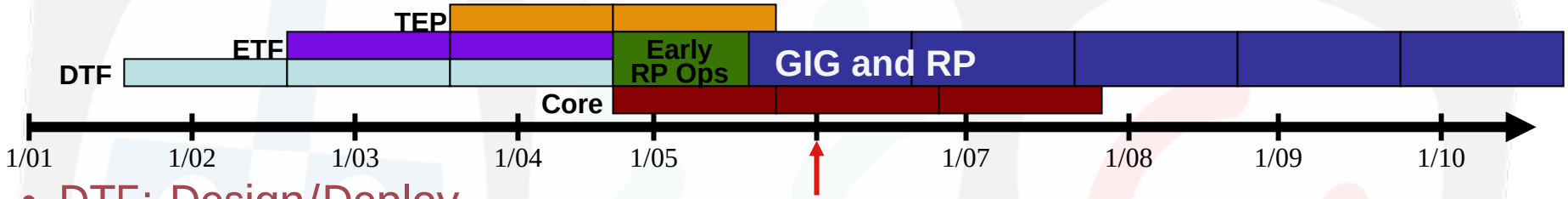
TeraGrid™

TeraGrid Objectives

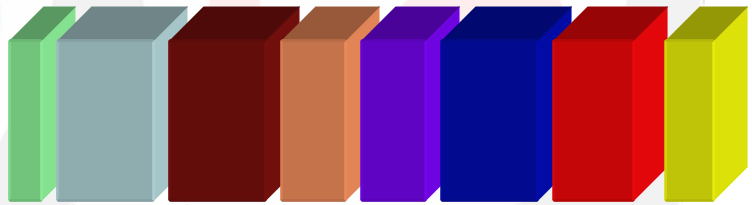
- DEEP Science: Enabling Terascale Science
  - Make Science More Productive through an integrated set of very-high capability resources.
- WIDE Impact: Empowering communities
  - Bring TeraGrid capabilities to the broad science community.
- OPEN Infrastructure, Open Partnership
  - Provide a coordinated, general purpose, reliable set of services and resources.

TeraGrid

TeraGrid Timeline



- DTF: Design/Deploy
  - UC/ANL, Caltech, NCSA, SDSC
  - Intel 64bit - Homogeneous
- ETF: Add Heterogeneity
  - +PSC
  - Add IBM SMP, HP
- TEP: Expand
  - + TACC, PU, IU, ORNL
  - + IA32
- FY05-09: Operations & Science Outreach
  - FY05-6 Add XT3, SGI SMP, BG/L
  - 8 RP Sites
  - Grid Infrastructure Group



Resource Providers

Networking, Operations & Security	Community Engagement: Science Gateways
Software Integration	User Support Coordination

Integration Team (Grid Infrastructure Group)

TeraGrid

TeraGrid Resources

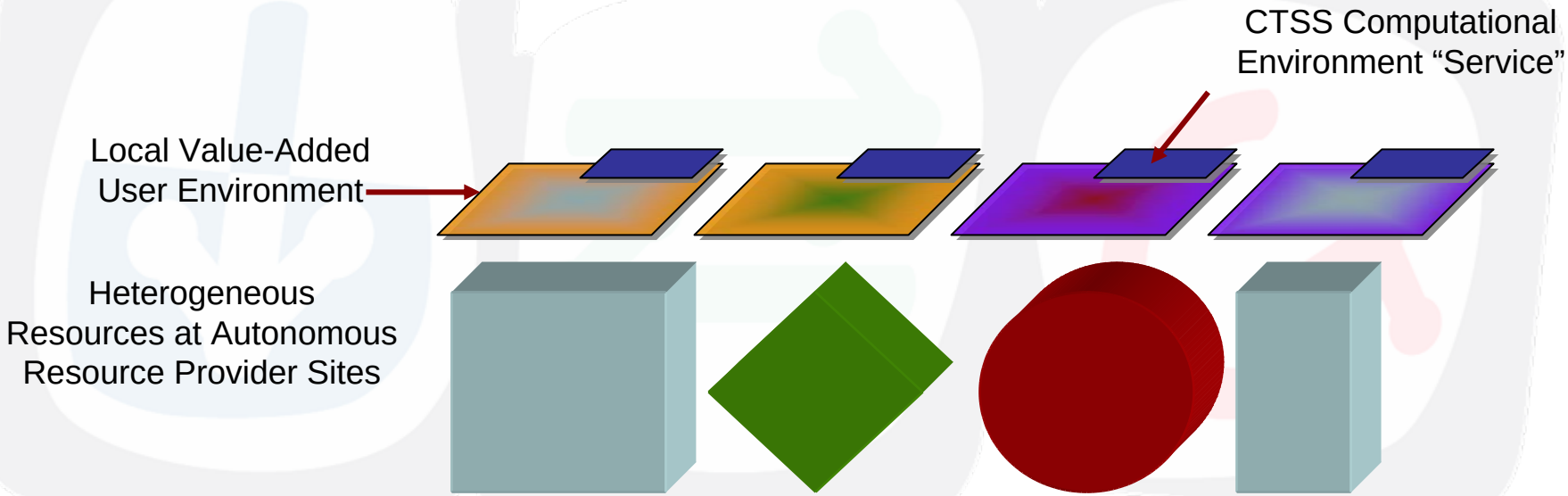
	ANL/UC	IU	NCSA	ORNL	PSC	Purdue	SDSC	TACC
Computational Resources	Itanium 2 (0.5 TF) IA-32 (0.5 TF)	Itanium2 (0.2 TF) IA-32 (2.0 TF)	Itanium2 (10.7 TF) SGI SMP (7.0 TF) Dell Xeon (17.2TF) IBM p690 (2TF) Condor Flock (1.1TF)	IA-32 (0.3 TF)	XT3 (10 TF) TCS (6 TF) Marvel (0.3 TF)	Hetero (1.7 TF) IA-32 (11 TF) *Opportunistic*	Itanium2 (4.4 TF) Power4+ (15.6 TF) Blue Gene (5.7 TF)	IA-32 (6.3 TF)
Online Storage	20 TB	32 TB	1140 TB	1 TB	300 TB	26 TB	1400 TB	50 TB
Mass Storage		1.2 PB	5 PB		2.4 PB	1.3 PB	6 PB	2 PB
Net Gb/s, Hub	30 CHI	10 CHI	30 CHI	10 ATL	30 CHI	10 CHI	40 LA	10 CHI
Data Collections # collections Approx total size Access methods		5 Col. >3.7 TB URL/DB/ GridFTP	> 30 Col. URL/SRB/DB/ GridFTP			4 Col. 7 TB SRB/Portal/ OPeNDAP	>70 Col. >1 PB GFS/SRB/ DB/GridFTP	4 Col. 2.35 TB SRB/Web Services/ URL
Instruments		Proteomics X-ray Cryst.		SNS and HFIR Facilities				
Visualization Resources RI: Remote Interact RB: Remote Batch RC: RI/Collab	RI, RC, RB IA-32, 96 GeForce 6600GT		RB SGI Prism, 32 graphics pipes; IA-32		RI, RB IA-32 + Quadro4 980 XGL	RB IA-32, 48 Nodes	RB	RI, RC, RB UltraSPARC IV, 512GB SMP, 16 gfx cards

100+ TF
8 distinct architectures

3 PB Online Disk

>100 data collections

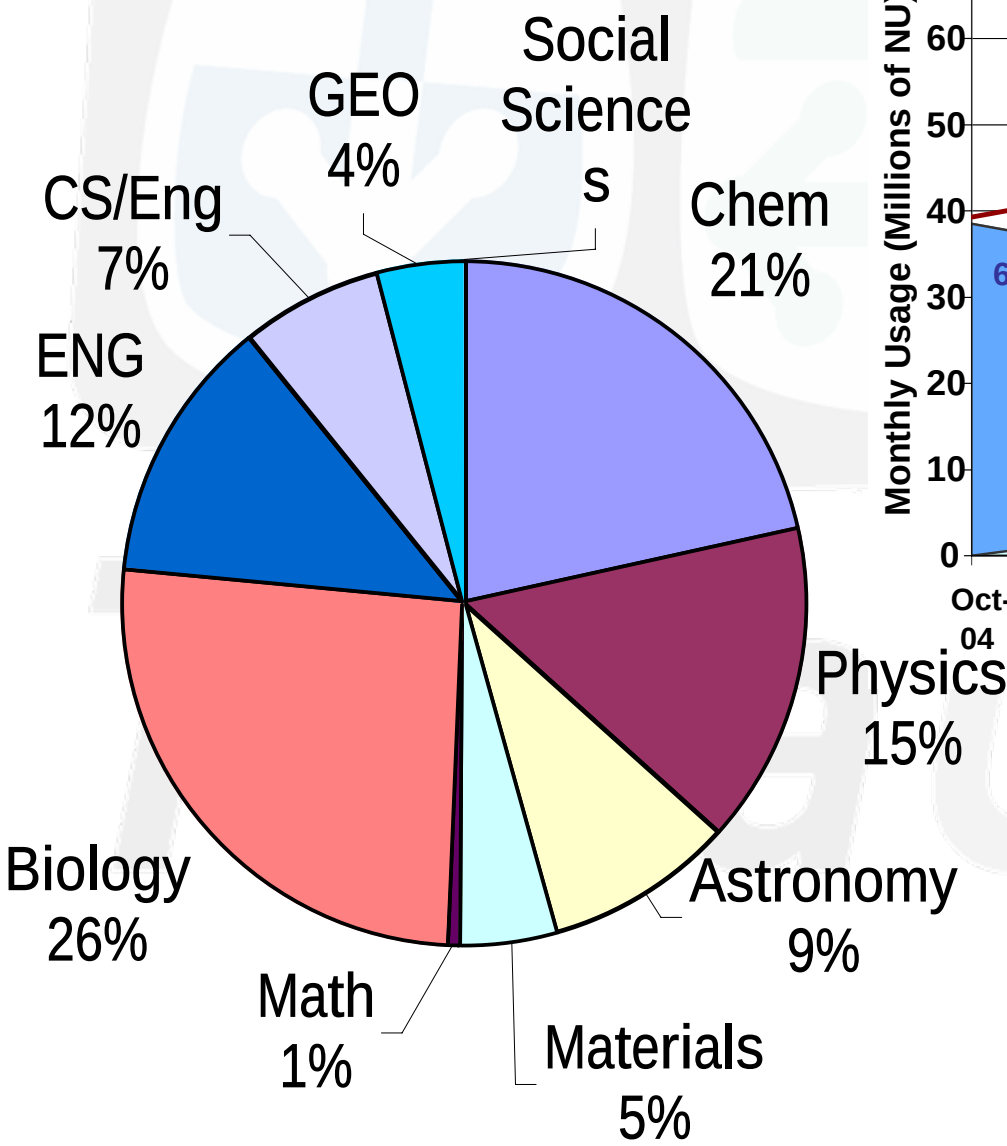
TeraGrid Architecture

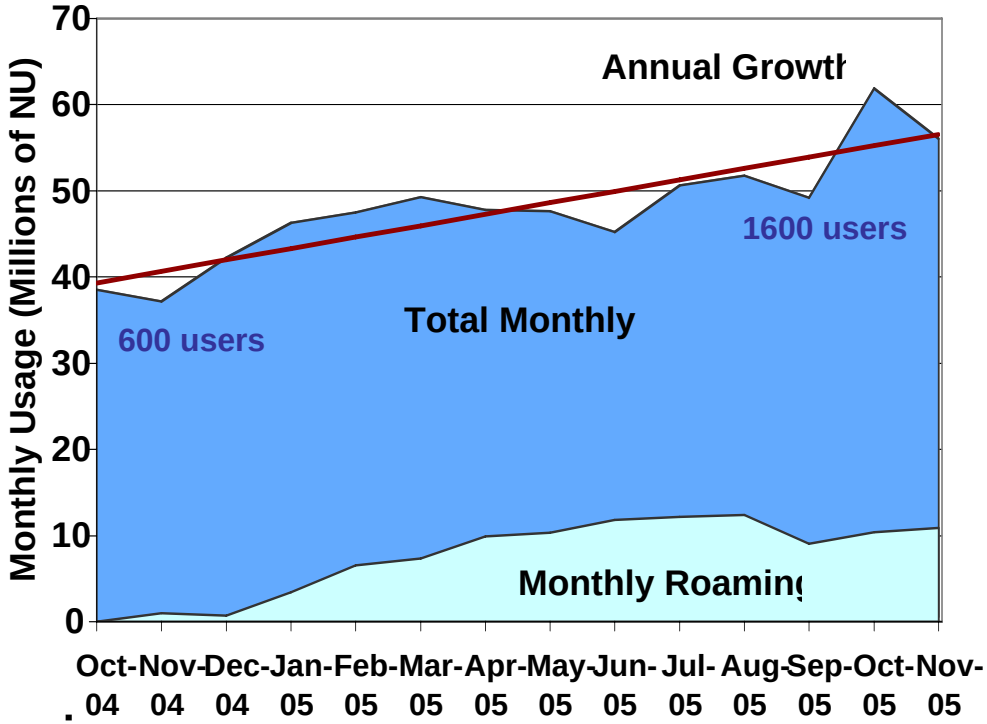


- **A single point of contact** for user assistance.
- **A common allocation process** that includes a currency usable on all systems, while preserving the need to provide specific machine access to users with specific needs.
- **A common access service and environment** on all platforms, allowing users to readily move from machine to machine as needed. *Learn Once; Run Anywhere*.
- **Services to assist users in harnessing the right TeraGrid platforms for each part of their work**, ranging from tightly-coupled applications (MPICH-G2) to workflow (Condor-G, GridShell), file staging (GridFTP/RFT) and remote file I/O (0.5 PB GPFS WAN filesystem), supported by common authentication (GSI), and in 2006 Web services via GT4.
- **New capabilities driven** by tight feedback loop with **users** via surveys and hands-on projects.
- **Science Gateways** build on this architecture (common definitions, interfaces) to reach communities.

TeraGrid

TeraGrid Use



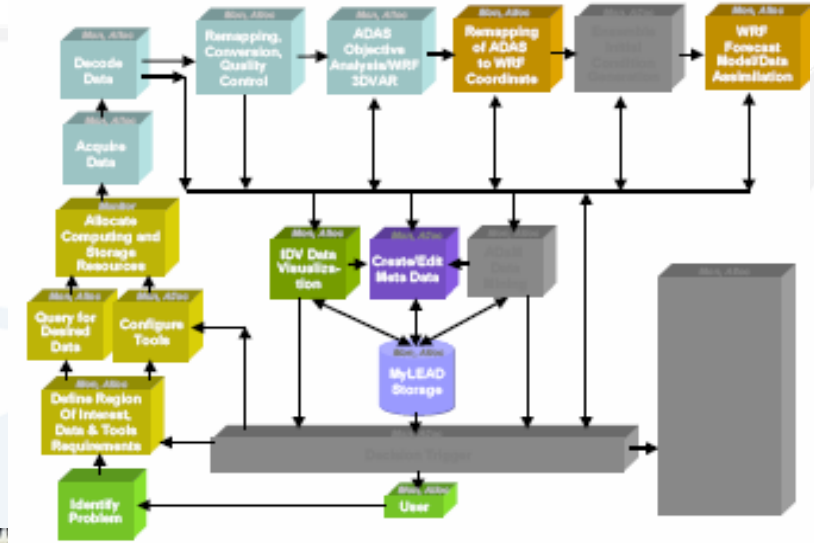


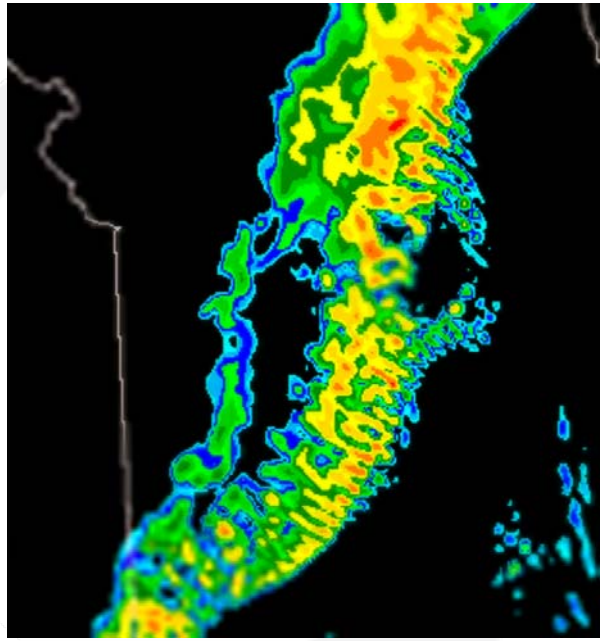
TeraGrid

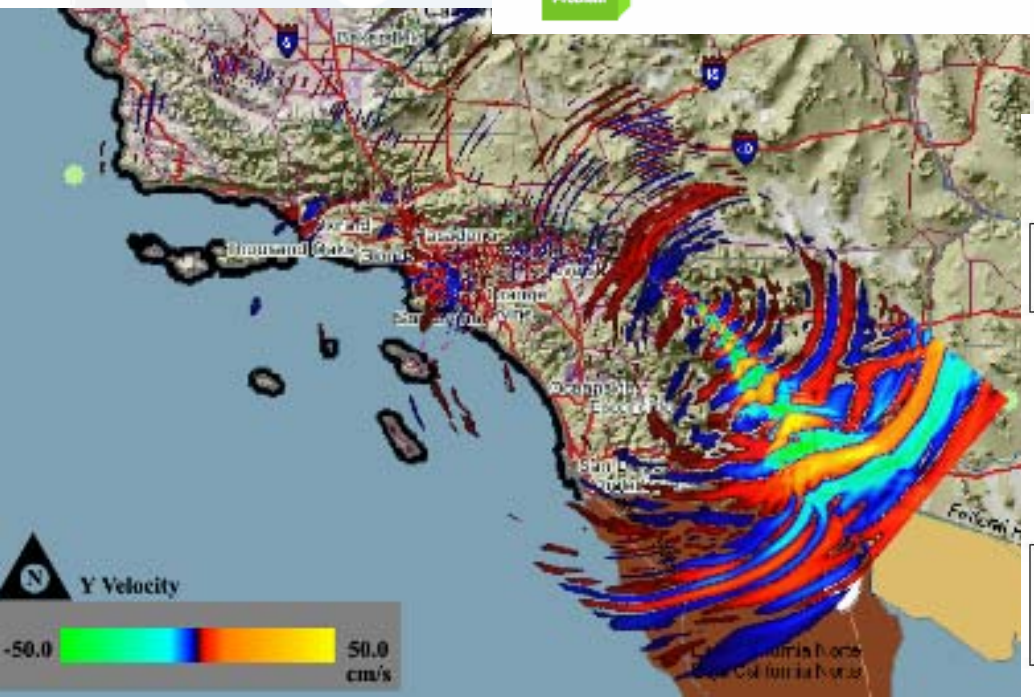
On Demand:

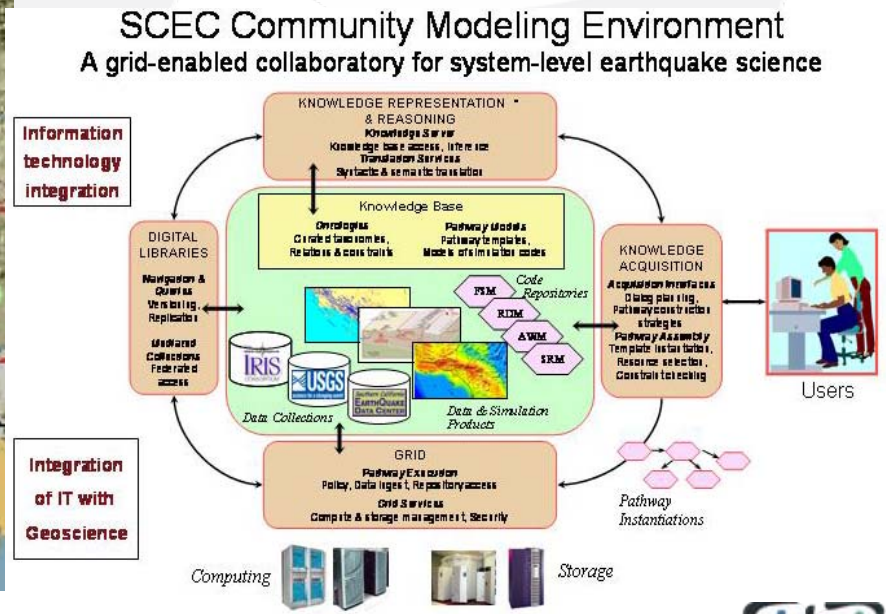
Predicting Severe Weather

Droegemeier (OU) and LEAD





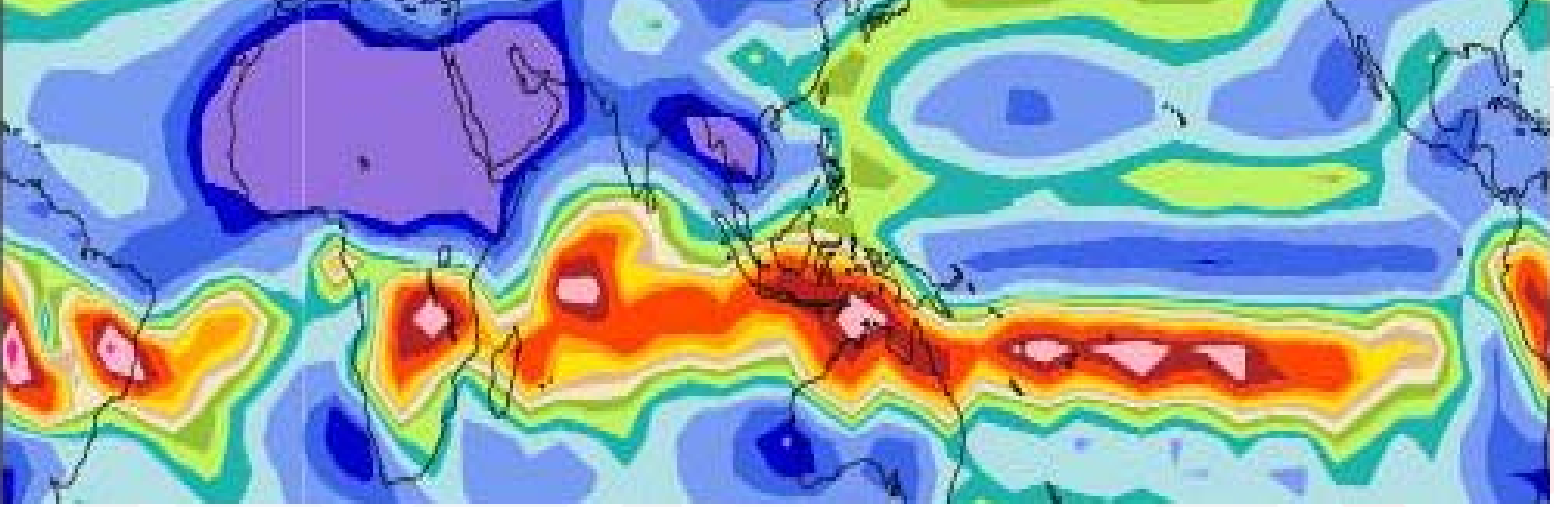




Large Data; Virtualized Resources: Earthquake Simulation

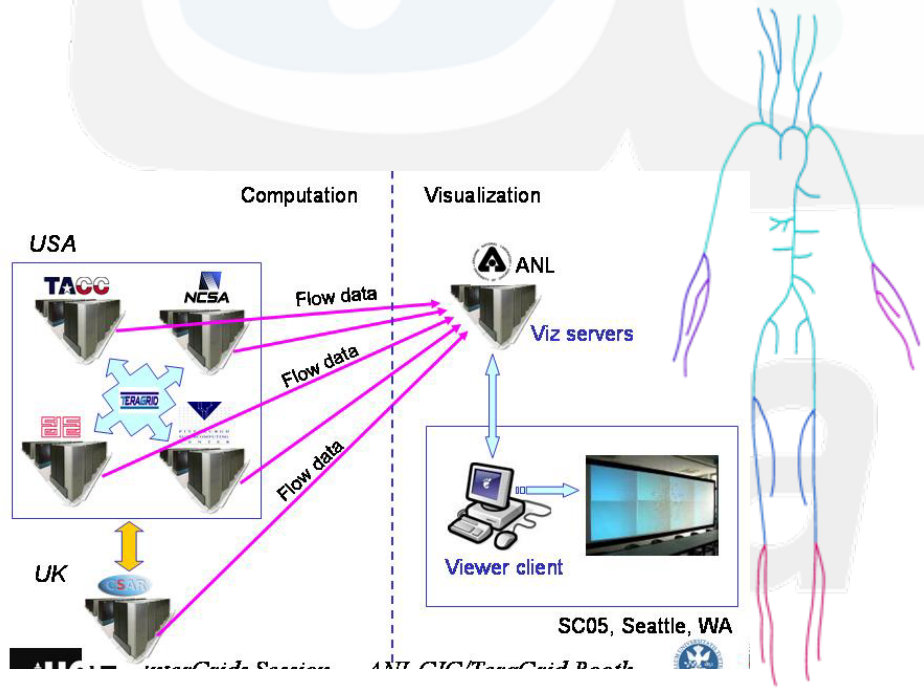
Olsen (SDSU), Okaya (USC), Southern California Earthquake Center

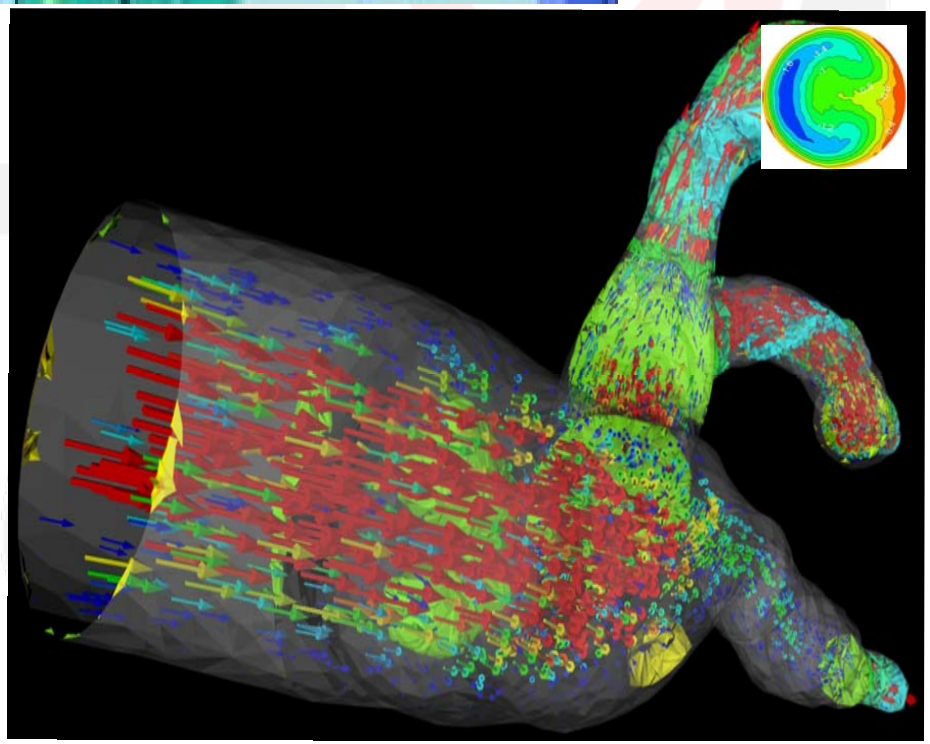
TeraGrid



Virtualized Resources, Ensembles: FOAM Climate Model

Liu (UWisc)



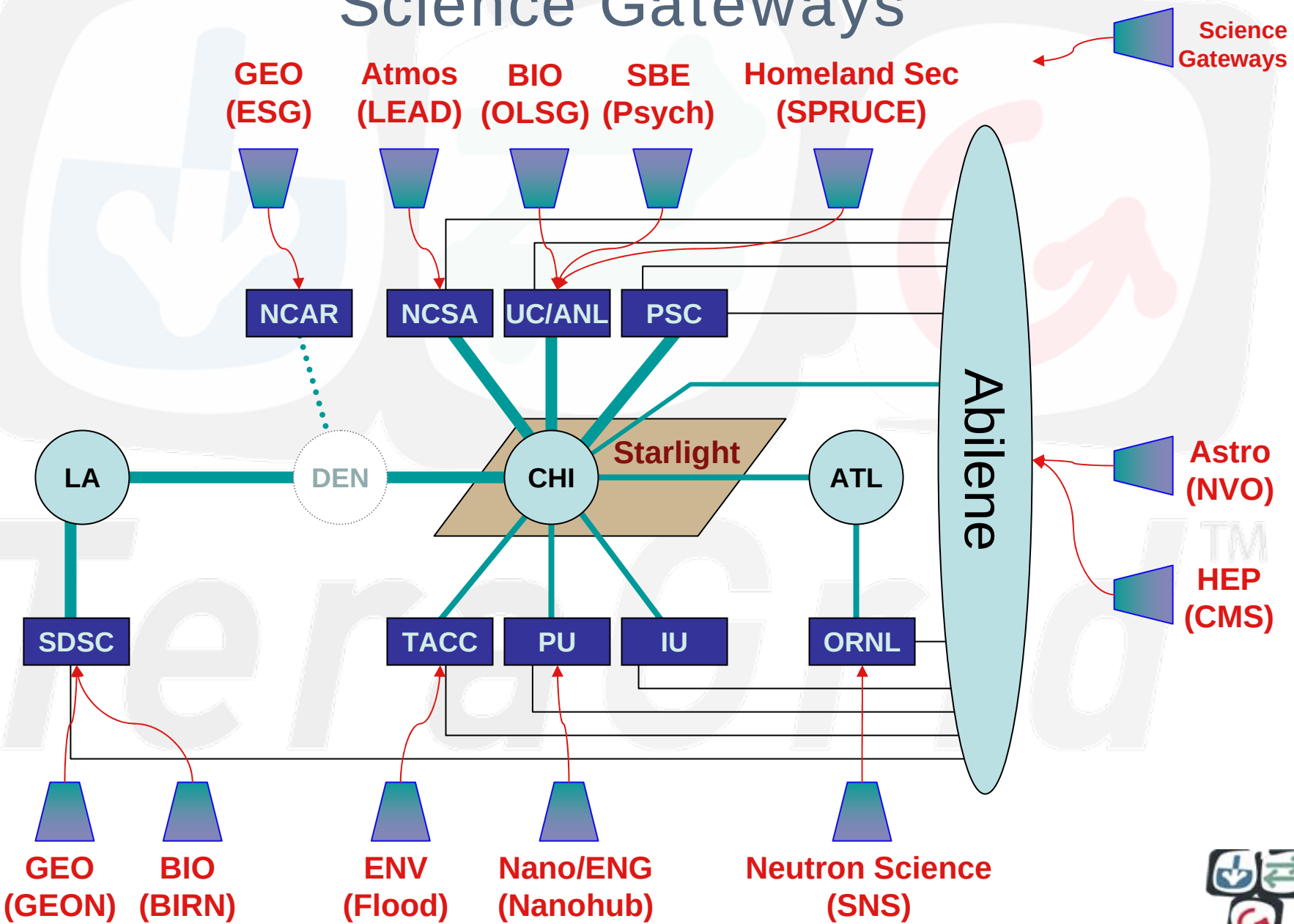


Coupled Simulation: Full Body Arterial Tree Simulation

Karniadakis (Brown)

TeraGrid

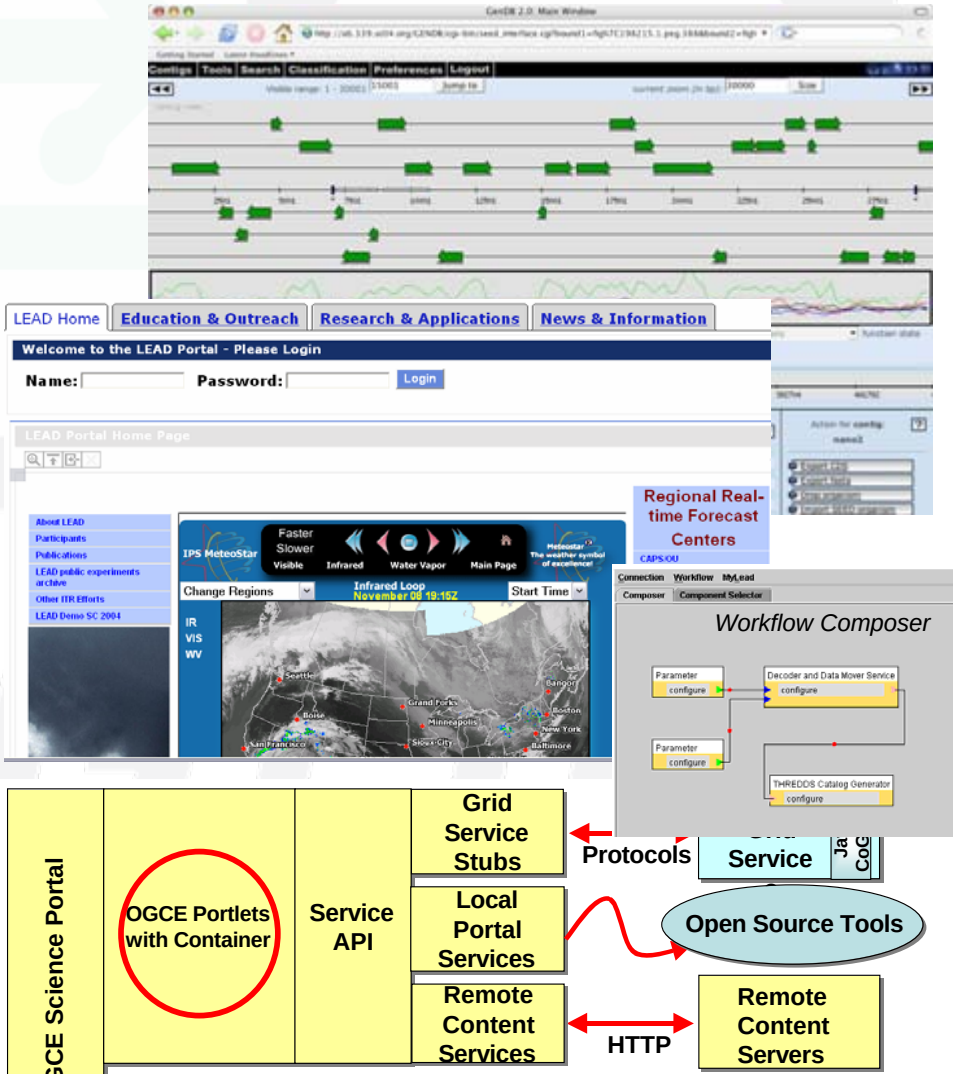
Science Gateways



TeraGrid

Science Gateways
A new initiative for the TeraGrid

- Increasing investment by communities in their own cyberinfrastructure, but heterogeneous:
  - Resources
  - Users – from expert to K-12
  - Software stacks, policies
- Science Gateways
  - Provide "TeraGrid Inside" capabilities
  - Leverage community investment
- Three common forms:
  - Web-based Portals
  - Application programs running on users' machines but accessing services in TeraGrid
  - Coordinated access points enabling users to move seamlessly between TeraGrid ar other grids.



TeraGrid

Building an Open Life Science Gateway

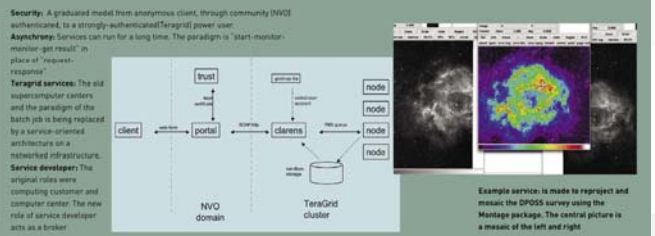
- Identifying life science communities that can be empowered by TeraGrid
  - Initial target ⇒ NIH Bioinformatics Resource Centers
    - 8 centers created to serving the pathogen research community
    - Leveraging NIH NIAID resources
- Develop web services/grid services based interfaces to existing compute intensive tools
  - Targeting key applications in genomics analysis, drug target analysis, computational molecular biology, cell biology and molecular evolution
- Create service bundles that can be deployed on TeraGrid Resources as persistent services (services and needed datasets)
  - Clusters of related application back-ends and required infrastructure
- Integrate web services/compute back-ends to existing end-user tools
  - Web based tools, native desktop tools, embedded applications
- Tool and community based allocations
  - Using the TeraGrid via enabled tools will be transparent to the user

Science Gateway Examples

Nano-technology: nanoHUB.org

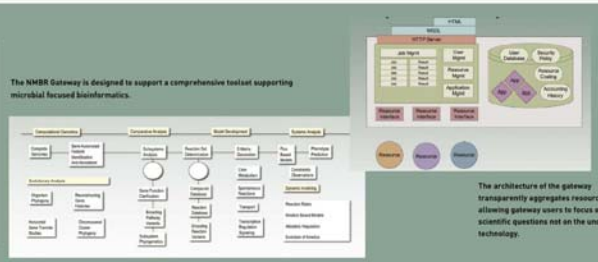
NCN PI: Mark Lundstrom,
Gerhard Klimeck, Purdue University
TeraGrid PI: Sebastien Goasguen,
Purdue University

NVO Service Framework



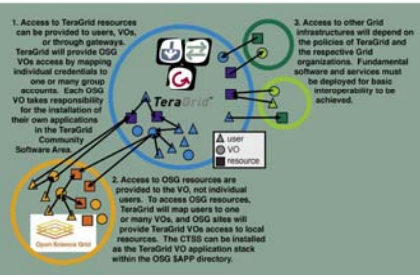
co-PI: Roy Williams, Caltech
co-PI: Julian Bunn, Caltech

National Microbial Bioinformatics Resource Center (NMBR)



PI: Rick Stevens, Argonne National Laboratory / University of Chicago
PI: Olaf Schneewind, University of Chicago

Open Science Grid



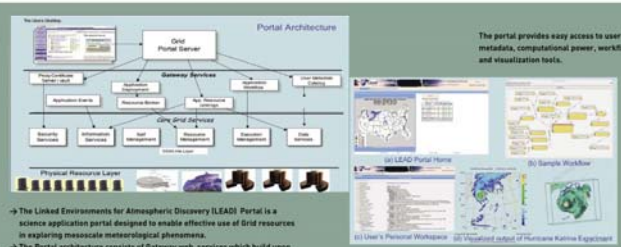
Bioinformatics Science Gateway

PI: Daniel A. Reed, RENCI

Flood Modeling

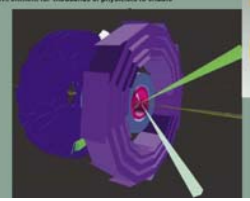
Project Lead: Bill Barth, TACC
Project Members: David Guzman, Patrick Hurley Tomislav Urban, TACC

LEAD Gateway



PI: Dennis Gannon, Indiana University

Grid Portals for LHC Particle Physics



Co-PIs: Julian Bunn & Roy Williams, California Institute of Technology Booth 428

As well as additional gateway projects that have joined us or are planning to join, including… University of Buffalo, BIRN ,NEES, GEON, Several NCAR projects, Cornell (large data collections), LSU (coastal modeling), IU Hydra Portal

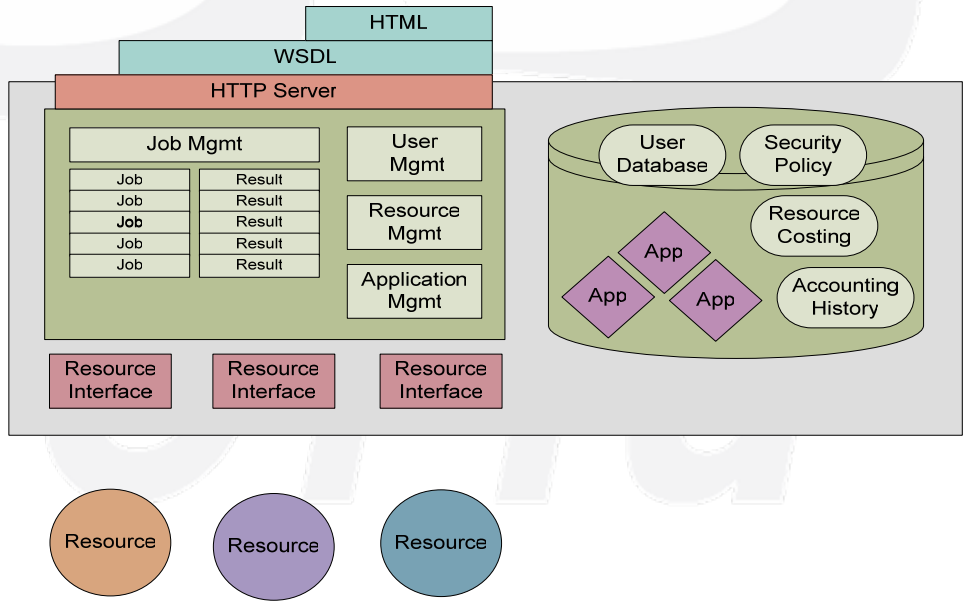
TeraGrid

Life Science Gateway Architecture

The Life Science Gateway provides WS based access to bioinformatics resources – databases, computing grids, and applications.

Gateway Resources:
- Web Services
- Tutorials
- Applications
- Data Repositories
- “Mirrors” for External Service Interfaces

Gateway Resources Access:
- Web Pages
- Web Services
- Applications
- Data Repository



TeraGrid

Planned LSGW Resources

- Genomics
  - Genome Assembly
  - Gene Calling
  - Sequence Analysis
  - Domain Analysis
  - Functional Coupling
- NCBI Genome Data
- BRC
- KEGG
- Uniprot
- SEED

- Phylogeny
  - Sequence Alignment
  - Tree Building
  - HGT Search
- TIGR Tools
- NCBI Tools
- Hmmer, Pfam
- InterProScan
- ScopMap
- SEED

LSGW Web Services

The Life Science Gateway supports web services access to the applications, tools, and data it hosts.

Web Service Toolkits:

SOAP::Lite

Python's ZSI

Microsoft .Net

Apache Axis

gSOAP

Workflow Tools:

Taverna

Keppler

Apple's Automator

Example SOAP::Lite Program:

```
#!/usr/bin/perl -w
use SOAP::Lite;
my $service = SOAP::Lite
       -> service('http://lsgw.mcs.anl.gov/gw/wsdl/SquareService.wsdl');
my $result = $service->getSquare(2);
print $result, "\n";
```

TeraGrid

LSGW Computational Scheduling

The Life Science Gateway internally manages job requests, application deployments, data transfer, job dispatching and job accounting.

Gateway resources:

- Can be added and removed
- Are intelligently managed
  - Small jobs can be executed on small resources
  - Large expensive resources can be allocated to users based on gateway policies
- Have interfaces for retrieving accounting and audit information
  - Resources can ask to have jobs from certain users disabled
  - Resources can ask the Gateway for information about the user who submitted a specific job

The typical set of resource the gateway will utilize include:

- Host Local Execution – Runs jobs on the gateway
- Campus Local Execution – Runs jobs on resources local to the user
- TeraGrid Execution

Community Based Allocations

- Significant “Grid” resources are allocated directly to a proxy for a community, examples:
  - Proxy representing a TG Science gateway team
  - PI or proxy for a NIH Research Resource
  - PI of proxy for a NSF S+T Center
- The proxy’s responsibility is to insure the community is exploiting the TeraGrid and to adjust mix of applications and tools in real time to maximize the scientific output
- Tracking and accounting is done by community as a group not by specific users or tools

Applications Based Allocations

- Significant “Grid” resources are allocated directly to a proxy for one or more applications or tools
  - ScopMap, InterProScan, HMMR, Blast, etc.
- The proxy’s responsibility is to insure that the latest version of the tool is available on many platforms and via many modalities
  - Web services, portals etc.
- Tracking, and accounting is done by application (not by user or problem, cross community use is ok)